استناد به این مقاله: هاشم‌زاده، محمدجواد؛ نخعی، زینب؛ مرادی مقدم، حسین (۱۳۹۲). کاربرد و تعدیل قانون زیف و الگوی بازو در بازشناسی واژه‌های بازدارنده زبان فارسی با استفاده از خوشه زبانی مقالات علمی – پژوهشی رشته کتابداری و اطلاع‌رسانی. پژوهش‌نامه کتابداری و اطلاع‌رسانی،۳ (۲)، ۲۰۸-۱۹۱.

کاربرد و تعدیل قانون زیف و الگوی آماری زو در بازشناسی واژه‌های بازدارنده زبان فارسی با استفاده از خوشه زبانی مقالات علمی- پژوهشی رشته کتابداری و اطلاع‌رسانی

دکترمحمد جواد هاشم زاده[۱] ، زینب نخعی[۲]، حسین مرادی مقدم[۳]

دریافت: ۱۳۹۱/۷/۲۹ پذیرش: ۱۳۹۲/۴/۴

چکیده

هدف: *شناسایی و استخراج سیاهه‌هایی نظام‌مند از واژه‌های بازدارنده به منظور استفاده در نمایه‌سازی خودکار متن‌های فارسی رشته کتابداری و اطلاع‌رسانی*

روش: *روش تحلیل محتوا است. جامعه پژوهش، ۵۶ مقاله بودند که ۲۰ مقاله با روش نمونه‌گیری تصادفی ساده انتخاب شدند.*

یافته‌ها: *از مجموع ۱۵۵۵۷ واژه موجود در متن مقالات، مطابق با الگوی زو، قبل از تعدیل واژه‌ها، ۱۳۶۸ و بعد از تعدیل، ۴۶۸ واژه؛ مطابق قانون زیف نیز قبل از تعدیل، ۲۱۷ و بعد از تعدیل، ۶۰۷ واژه به عنوان واژه بازدارنده شناخته شدند. هم‌چنین از مجموع ۱۹۸۹ واژه موجود در چکیده مقالات، مطابق با الگوی زو قبل از تعدیل واژه‌ها، ۱۴۸ و بعد از تعدیل، ۱۷۳ واژه و بر اساس قانون زیف، قبل از تعدیل ۶۰ و بعد از تعدیل، ۱۸۶ واژه به عنوان واژه بازدارنده استخراج شدند. در هر دو روش رابطه مستقیمی بین بسامد واژه‌ها و احتمال بازدارنده بودن آن‌ها مشاهده شد. بالاترین درصد واژه‌های بازدارنده (۳۹/۴۴ درصد) بدون احتساب بسامد، در متن مقالات و با کاربرد الگوی آماری زو به دست آمد. نتایج این پژوهش به افزایش کارایی، کاهش حجم فایل درونداد و صرفه‌جویی در زمان و هزینه ذخیره و بازیابی اطلاعات منجر می‌شود.*

کلیدواژه‌ها: *واژه‌های بازدارنده؛ بسامد واژگانی؛ قانون زیف؛ نمایه‌سازی خودکار؛ الگوی آماری زو.*

---

۱ . استادیار گروه کتابداری و اطلاع‌رسانی دانشگاه بیرجند؛ hashemzadeh@birjand.ac.ir

۲ . کارشناس ارشد کتابداری و اطلاع‌رسانی دانشگاه بیرجند؛ z.nakhaie@yahoo.com

[۳] مدرس دانشگاه بیرجند، moradymoghadam@gmail.com

مقدمه

در طول چند دهه گذشته و با تلاش‌های متخصصان اطلاع‌رسانی، نمایه‌سازی خودکار توانسته است به عنوان ابزاری برای سازماندهی و در دسترس قرار دادن حجم عظیم اطلاعات موجود یا به عبارتی مقابله با انفجار اطلاعات مطرح گردد که در آن کل فرایند نمایه‌سازی، اعم از استخراج کلیدواژه‌ها، مرتب کردن مدخل‌ها و... توسط رایانه انجام می‌گیرد. در این نوع نمایه‌سازی به روش‌هایی که بر فراوانی نسبی کلمات در متن مبتنی هستند، تکیه می‌شود. تعیین واحدهای متنی و مشخص کردن حد و مرز واژه برای ماشین از مسائل اساسی در گزینش اصطلاحات نمـایه‌ای در نمـایه‌سـازی خودکار است (گیلوری، ۱۳۷۹). به علاوه، امکان تشخیص واژه‌های مفهومی از واژه‌های بازدارنده، در فرایند انتخاب اصطلاحات نمایه‌ای تأثیر بسزایی دارد. آنچه مسلم است، ماشین این امکان تشخیص را تنها از طریق تطبیق واژه‌های استخراج شده از متن یا منتسب شده به متن با فهرستی که واژه‌های غیرمجاز (بازدارنده)[1] نامیده می‌شـود، به دسـت می‌آورد (سنجی، ۱۳۸۷). البته تعیین واژه‌های غیرمجاز کار راحتی نیست، زیرا هر واژه ممکن است در نظامی مجاز و در نظامی دیگر غیرمجاز تلقی شود (تیلور[2]،۱۳۸۱).

در بازیابی اطلاعات عموماً کلماتی را که به صورت مکرر در متن ظاهر می‌شوند اما به تنهایی بار معنایی ندارند و در ارتباط با واژه‌های دیگر معنا می‌یابند و فقط به دلیل دستوری مورد استفاده قرار می‌گیرند، واژه‌های غیرمجاز می‌نامند. شناسایی این‌گونه از واژه‌ها یکی از مراحل اساسی در امر سازماندهی، ذخیره و بازیابی اطلاعات است که کاربرد آن در رشته‌های علمی و در فرایندهای نمایه‌سازی از قبیل نمایه‌های کوئیک مشخص شده است. به علاوه، سیاهه بازدارنده یک ابزار بنیادی و ضروری برای خوشه‌بندی مدرک و تحلیل شباهت مدارک و سایر اموری که با تحلیل مدارک سروکار دارند، می‌باشد که استفاده از آن منجر به کارایی بهتر این امور می‌شود. در صورتی که این واژه‌ها قبل از فرایند نمایه‌سازی مدارک مشخص و فهرست آن‌ها برای کنترل به رایانه داده شود، باعث صرفه‌جویی در زمان و حجم بایگانی‌های نمایه شده و کارایی فرایند نمایه‌سازی را بین ۳۰ تا ۵۰ درصد افزایش می‌دهد؛ هم‌چنین به میزان زیادی از بازیابی مدارک نامرتبط و ریزش کاذب در جست‌وجو جلوگیری خواهد کرد (زو[3] و دیگران، ۲۰۰۶). اما با وجود این، قواعد روشنی برای ایجاد چنین فهرستی وجود ندارد و بیشتر پژوهشگران از روش‌های مختلف برای استخراج واژه‌های بازدارنده استفاده می‌کنند مانند: ۱. بسامد[4]؛ ۲. کارکرد

1. Stop Words
2 .Taylor
3. Zou
4.Frequency

نحوی؛ ۳. کاربرد قانون زیف[1]؛ ۴. محاسبه نمره واژه و ۵. محاسبه آنتروپی[2] واژه. با توجه به این که اغلب شیوه‌های شناسایی و استخراج واژه‌های بازدارنده بر بسامد واژگانی به عنوان بهترین و متداول‌ترین راه تأکید دارند و هم‌چنین بر اساس قانون زیف و الگوی آماری زو و همکاران (۲۰۰۶) که مشخص ساخت واژه‌های دارای بسامد بالا و توزیع ثابت در مدارک مختلف به عنوان واژه‌های بازدارنده در نظر گرفته می‌-شوند، پژوهش حاضر بر آن است که قابلیت به کارگیری دو روش مبتنی بر بسامد را در زبان فارسی مورد بررسی قرار دهد و بر این اساس واژه‌های بازدارنده موجود در متن و چکیده مقالات علمی- پژوهشی رشته کتابداری و اطلاع‌رسانی را شناسایی نماید.

پرسش‌های پژوهش

۱. واژه‌های بازدارنده موجود در متن مقاله‌ها جهت به کارگیری در نمایه‌سازی خودکار مدارک فارسی رشته کتابداری و اطلاع‌رسانی بر اساس الگوی آماری زو، کدام است؟

۲. واژه‌های بازدارنده موجود در چکیده مقاله‌ها جهت به کارگیری در نمایه‌سازی خودکار مدارک فارسی رشته کتابداری و اطلاع‌رسانی بر اساس الگوی آماری زو، کدام است؟

۳. واژه‌های بازدارنده موجود در متن مقاله‌ها جهت به کارگیری در نمایه‌سازی خودکار مدارک فارسی رشته کتابداری و اطلاع‌رسانی بر اساس قانون زیف، کدام است؟

۴. واژه‌های بازدارنده موجود در چکیده مقاله‌ها جهت به کارگیری در نمایه‌سازی خودکار مدارک فارسی رشته کتابداری و اطلاع‌رسانی بر اساس قانون زیف، کدام است؟

۵. تا چه حد بین سیاهه‌های واژه‌های بازدارنده به دست آمده از متن و چکیده، در الگوی زو هم‌خوانی وجود دارد؟

۶. تا چه حد بین سیاهه‌های واژه‌های بازدارنده به دست آمده از متن و چکیده، در روش زیف هم‌خوانی وجود دارد؟

۷. تا چه حد بین سیاهه‌های واژه‌های بازدارنده به دست آمده از متن مقاله‌های دو روش زو و زیف هم‌خوانی وجود دارد؟

---

1. Zipf law

۲. آنتروپی یکی از مقیاس‌های بنیادی در نظریه اطلاعات است که میزان وضعیت تصادفی بودن یک علامت (سیگنال) یا یک اتفاق تصادفی را محاسبه می‌کند و یا میزان اطلاعاتی که توسط یک علامت حمل می‌شود را نشان می‌دهد.

۸. تا چه حد بین سیاهه‌های واژه‌های بازدارنده به دست آمده از چکیده مقاله‌های دو روش زو و زیف هم‌خوانی وجود دارد؟

۹. واژه‌های بازدارنده چه حجمی از متن و چکیده مقاله‌های حوزه مورد بررسی را تشکیل می‌دهند؟

پیشینه پژوهش

از آن‌جا که واژه‌های بازدارنده نقش مهمی در بازیابی اطلاعات در پایگاه‌های اطلاعاتی دارند و حذف آن‌ها در نمایه‌سازی خودکار سبب سرعت بخشیدن به پردازش اطلاعات و در نتیجه صرفه‌جویی در زمان و فضای ذخیره‌سازی می‌شود، پژوهش‌های زیادی در سراسر جهان و در زبان‌های مختلف پیرامون این موضوع انجام شده است.

فاکس[1] (۱۹۹۰) از اولین افرادی بود که فهرست واژه‌های بازدارنده را تهیه کرد. وی از روش بسامد استفاده کرد و پربسامدترین واژه‌های به دست آمده از پیکره زبانی براون[2] را با در نظر گرفتن نقطه برش[3] بسامد، به عنوان واژه‌های بازدارنده عمومی در زبان انگلیسی پیشنهاد کرد. فهرست او شامل ۴۲۱ واژه بازدارنده بود که بعدها این فهرست در سیستم بازیابی اوکاپی[4] مورد استفاده قرار گرفت.

ویلبور و سیروتکین[5] (۱۹۹۲) مجموعه‌ای از ۷۱۳۱۱ مدرک مدلاین در حوزه بیوتکنولوژی را بررسی قرار دادند و دریافتند که با حذف واژه‌های بازدارنده، حجم واژه‌های مدارک حدود ۷۵ درصد کاهش می‌یابد.

لازارینیس[6] (۲۰۰۷) مراحل ایجاد یک سیاهه بازدارنده برای زبان یونانی و تأثیر حذف آن از پرسش‌های کاربران را مورد بررسی قرار داد و دریافت زمانی که واژه‌های بازدارنده از پرسش کاربران حذف شدند، جست‌وجو سریع‌تر انجام شد و تعداد صفحات مرتبط بیش‌تری بازیابی شد.

در پژوهشی پاندی و سیدیکوئی[7] (۲۰۰۷) به ارزیابی تأثیر ریشه‌یابی و حذف کلمات بازدارنده بر بازیابی متون هندی پرداختند و به این نتیجه رسیدند که حذف کلمات بازدارنده به طور معنادار باعث

1.Fox

۲. این پیکره شامل ۵۰۰ مدرک بود که همه آن‌ها برای اولین بار در سال ۱۹۶۱ منتشر شدند. این پیکره تقریباً ۱۰۱۴۰۰۰ واژه را از متون جاری ۱۶۷ حوزه نوشتاری در بر داشت.

3 Cutt off Point
4 okapi
5 Wilbur & Sirotkin
6 Lazarinis
7 Pandey & Siddiqui

افزایش دقت بازیابی و کاهش حجم فایل نمایه شد.

در پژوهشی دیگر هائو و هائو[1] (۲۰۰۸) یک رویکرد خودکار برای ایجاد فهرست واژه‌های بازدارنده در رده‌بندی متون چینی ارائه نمودند و هر واژه‌ای که دارای بسامد بالا در مدرک و همبستگی آماری کم با سایر طبقه‌های موجود در طرح رده‌بندی متون چینی بود، به عنوان واژه بازدارنده انتخاب نمودند. آن‌ها به این نتیجه رسیدند که سیاهه آن‌ها که در رده‌بندی نوشته‌های چینی مؤثر است و باعث افزایش دقت و کاهش زمان رده‌بندی می‌شود.

در زبان فارسی سنجی (۱۳۸۷) پس از تعیین بسامد و نوع دستوری واژه‌های موجود در مقالات رشته‌های روانشناسی، علوم تربیتی و کتابداری و اطلاع‌رسانی، تعداد ۹۷۲۸۰ واژه (۱۲۹۱ واژه بدون احتساب بسامد) را به عنوان واژه‌های غیرمفهومی در سه رشته مورد مطالعه شناسایی نمود. هم‌چنین نشان داد که افعال، قیود، ضمایر، حروف، اصوات، اعداد و علائم سجاوندی به عنوان واژه‌های نمایه‌ای ظاهر نمی‌شوند.

در پژوهشی دیگر داورپناه، سنجی و آرمیده (۲۰۰۹) طی دو مرحله به بررسی ۶۳ مقاله در حوزه-های روان‌شناسی، علوم تربیتی و کتابداری و اطلاع‌رسانی و همچنین پیکره ایجاد شده به وسیله روزنامه همشهری پرداختند. در مرحله اول ۷۴۶ واژه و در مرحله دوم ۴۲۲ واژه را به عنوان واژه بازدارنده معرفی کردند. به طور کلی از مرور پیشینه پژوهش می‌توان دریافت که پژوهش‌های بسیاری در مورد واژه‌های بازدارنده (ساووی[2]، ۱۹۹۹)؛ (ابوالخیر[3]،۲۰۰۶)؛ بسامد (ادموندسون و وایلز[4]،۱۹۵۹)؛ (برگ[5]، ۱۹۹۷) و قانون زیف (فرانسیس و کوسرا[6]، ۱۹۶۷)؛ (آراپوو و فیمووا و اشرایدر[7]، ۱۹۷۵) به ویژه در خارج کشور و پژوهش‌های کم‌تری در مورد زبان فارسی صورت گرفته است اما بر اساس بررسی‌های انجام شده پژوهشی مقایسه‌ای در مورد شناسایی واژه‌های بازدارنده مطابق با الگوی آماری زو و همکاران (۲۰۰۶) و قانون زیف یافت نشد. در بیش‌تر این پژوهش‌ها از روش بسامد واژگانی استفاده شده است و اکثر فهرست‌های بازدارنده به دست آمده، کاربرد عمومی دارند و فقط تعداد انگشت شمار و محدودی از این فهرست‌ها در حوزه‌های خاص موضوعی و به صورت تخصصی تهیه شده‌اند. محققان در پژوهش‌های خود از پیکره‌های گوناگونی استفاده کرده‌اند. تعدادی از آن‌ها پیکره زبانی خود را از میان پایگاه‌های اطلاعاتی یا وب‌سایت-

1. Hao & Hao
2. Savoy
3 .Abu-El Khair
4 . Edmundson & Wylls
5. Berg
6. Francis & kucera
7. Arapov, Fimova & Shreider

ها انتخاب نموده‌اند و عده‌ای دیگر مجموعه مدارک چاپی (مقاله، روزنامه، چکیده و مانند آن) را مورد سنجش قرار داده‌اند. با توجه به این که تاکنون پژوهشی مقایسه‌ای در مورد شناسایی واژه‌های بازدارنده از یک خوشه واژگانی با استفاده از الگوی تطبیقی مشاهده نشده است، ضرورت پرداختن به این مقوله اهمیت می‌یابد.

روش پژوهش

این پژوهش از نظر هدف کاربردی و به روش تحلیل محتوا انجام گرفته است. جامعه آماری این پژوهش ۵۶ مقاله چاپ شده در آخرین شماره منتشر شده سال ۱۳۸۹ مجلات علمی- پژوهشی کتابداری و اطلاع‌رسانی می‌باشد که از بین آن‌ها ۲۰ مقاله با روش نمونه‌گیری تصادفی ساده انتخاب شدند.

ابزار پژوهش

۱. تهیه خوشه زبانی و متن الکترونیکی مقاله‌ها و چکیده‌های آن‌ها: در این مرحله تنها متن مقاله‌ها بدون در نظر گرفتن پانویس‌ها، ارجاعات، فرمول‌ها، جداول، نمودارها، اعداد ریاضی و منابع و مآخذ در محیط نرم‌افزاری word تایپ شد تا شیوه تایپ آن‌ها یکدست شود (مانند جست‌وجو و جستجو).

۲. تعدیل دستوری واژه‌ها (تفکیک واژگان): بخش‌های یک واژه مطابق با معیارهای مطرح شده در پژوهش سنجی (۱۳۸۷) مورد استفاده قرار گرفت. در این معیارها دستورالعمل‌هایی در مورد صیغه‌ها و وجوه افعال، افعال مرکب، مصدرهای مرکب، اسامی مرکب، اسامی پیشوندی، میانوندی و پسوندی، گروه‌های حرف اضافه، اسمی، قیدی و هم‌چنین تفکیک واژه‌ها از یکدیگر بر اساس فاصله بین آن‌ها، ارائه گردیده است.

۳. اجرای الگوی زو/ قانون زیف:

الف) الگوی آماری زو: پس از تعیین بسامد تمام واژه‌های موجود در متن و چکیده مقالات مورد بررسی، مطابق فرمول، $P_{i,j}$ آن‌ها محاسبه شد. به این صورت که بسامد واژه در یک متن/ چکیده بر کل تعداد واژه‌های همان متن/ چکیده تقسیم شد. سپس با استفاده از دو نرم‌افزار SPSS و MINITAB برای هر واژه با توجه به فرمول‌های زیر، «میانگین احتمال[1]»، «واریانس احتمال[2]» و«ارزش آماری واژه یا ضریب

1. Mean of Probability
2. Variance of Probability

تغییرات[1]» در کلیه متن‌ها و چکیده‌های آن‌ها محاسبه شد:

جدول ۱. فرمول‌های مربوط به الگوی زو

میانگین احتمال	واریانس احتمال	ارزش آماری واژه
$MP(W_j)=\frac{\sum_{1\le i\le N} P_{i,j}}{N}$ N=تعداد مدارک است که در این پژوهش ۲۰ مقاله می‌باشد.	$VP(W_j)=\frac{\sum_{1\le i\le N}\left(P_{i,j}-\overline{P_{i,j}}\right)^2}{N}$ $\overline{P}$ میانگین P های یک واژه.	$SAT(W_j)=\frac{\sqrt{Vp(W_j)}}{MP(W_j)}$

با بررسی ارزش آماری واژه‌های به دست آمده از این روش، نقطه برش در متن مقالات ۳۰۰ و در چکیده مقالات ۲۶۲ در نظر گرفته شد تا به این صورت بتوان تعداد واژه‌های غیرمفهومی بیش‌تری را در فهرست بازدارنده قرار داد. هر واژه‌ای از متن یا چکیده که ارزش آماری آن کم‌تر از این دو نقطه برش بود، به عنوان واژه بازدارنده انتخاب شد.

ب) قانون زیف: بر این اساس، واژه‌ها به ترتیب بسامد از زیاد به کم مرتب شدند. بالاترین بسامد واژه-های موجود در هر مقاله تعیین شد و بر اساس این فرمول نقطه عطف آن مقاله مشخص شد: $T=\frac{-1+\sqrt{1+8Fi}}{2}$ (هویدا ۱۳۷۸).

F_i: بالاترین بسامد واژه‌های موجود در یک مقاله است که در طبقه اول جدول زیف قرار گرفته است به این ترتیب در پژوهش حاضر برای هر متن و چکیده مقاله یک نقطه عطف تعیین شد و کلماتی از آن‌ها که بسامدی بالاتر از این نقاط عطف داشتند، به عنوان بازدارنده استخراج شدند و در یک فهرست قرار گرفتند و واژه‌های بازدارنده تکراری حذف شدند.

۴. تعدیل محتوایی و توجه به معنادار بودن یا غیرمعنادار بودن واژه‌ها: در شیوه اول یا قبل از تعدیل، واژه‌های بازدارنده بر اساس ارزش آستانه‌ای یا نقاط عطف تعیین شده استخراج شدند. اما در شیوه دوم یا بعد از تعدیل، ابتدا فهرست واژه‌های بازدارنده استخراج شده (قبل از تعدیل) مورد بازبینی قرار گرفت و واژه‌های معنادار آن‌ها حذف شد و سپس تعدادی از واژه‌های غیرمفهومی که در فهرست نیامده بودند، به فهرست اضافه شدند. پژوهش بر اساس قواعد ارائه شده آماری که کاربرد غیر قابل انکاری در شناسایی واژگان غیرمفهومی دارند و هم‌چنین بر اساس قاعده و نظریه زیف در مورد اصل کمترین کوشش

1. Statistical Value of word (SAT)

که پژوهش‌های بسیار زیادی پیرامون آن انجام پذیرفته، بنا شده است. هم‌چنین برای غنا بخشیدن به این یافته‌ها در زبان فارسی سعی گردید از معیارهای مطرح شده در پژوهش سنجی به عنوان مبنایی برای شناسایی واژه‌های مفهومی و غیرمفهومی و تشخیص و جداسازی واژه‌های بازدارنده از غیر بازدارنده استفاده شود. به این صورت که پس از تعیین نوع دستوری واژه‌های استخراج شده، اگر واژه‌ای جزء یکی از گروه‌های دستوری مورد اشاره در پژوهش ایشان (افعال، قیود، ضمایر و...) بود، به عنوان بازدارنده و در غیر این صورت به صورت واژه مفهومی در نظر گرفته شد.

یافته‌های پژوهش

در مورد هر فهرست، فقط ۱۰ واژه بازدارنده ابتدای فهرست‌های به دست آمده ارائه شده و فهرست‌های کامل از طریق ارتباط با پژوهشگر قابل دسترس است[۱]. لازم به ذکر است که علائم سجاوندی نیز در این پژوهش مورد پردازش قرار گرفته‌اند اما از جداول حذف گردیده است.

۱. واژه‌های بازدارنده موجود در متن مقاله‌ها جهت به کارگیری در نمایه‌سازی خودکار مدارک فارسی رشته کتابداری و اطلاع‌رسانی بر اساس الگوی آماری زو ، کدام است؟

جدول ۲. فهرست واژه‌های بازدارنده متن مقاله‌ها بر اساس الگوی زو «قبل و بعد از تعدیل»

ردیف	واژه	قبل از تعدیل	بعد از تعدیل	میانگین احتمال	واریانس احتمال	ارزش آماری
۱	آشکال	✓	–	۰.۰۰۰۱۰۵	۰.۰۰۰۰۰۰۰۰۵۳۱۳۱۶	۲۱۹.۵۳
۲	ابتدا	✓	–	۰.۰۰۰۳	۰.۰۰۰۰۰۰۰۱۱۷	۱۱۳.۹۴
۳	ابزار	✓	–	۰.۰۰۰۵۶۵	۰.۰۰۰۰۰۰۰۳۵۴	۱۰۵.۳
۴	ابعاد	✓	–	۰.۰۰۰۲۹	۰.۰۰۰۰۰۰۰۵۰۲	۲۴۴.۳۲
۵	اثر	✓	–	۰.۰۰۰۶۱۵	۰.۰۰۰۰۰۰۲۲۹	۲۴۶.۲۷
۶	احتمالاً	✓	✓	۰.۰۰۰۰۵	۰.۰۰۰۰۰۰۰۰۱۶۳۱۵۸	۲۵۵.۴۷
۷	ارائه	✓	✓	۰.۰۰۱۲۴۵	۰.۰۰۰۰۰۰۳۴۹	۱۴۹.۹۸
۸	ارائه شده	✓	✓	۰.۰۰۰۶۳۳	۰.۰۰۰۰۰۰۰۵۵۹	۱۱۸.۰۳
۹	ارتباط	✓	–	۰.۰۰۱۱۲۵	۰.۰۰۰۰۰۰۱۳۶	۱۰۳.۷۸
۱۰	اثرگذار	✓	–	۰.۰۰۰۰۳	۰.۰۰۰۰۰۰۰۰۰۴۳۱۵۷۹	۲۱۸.۹۸
	جمع	۱۳۶۸	۴۶۸			

۱. در صورت لزوم با آدرس ایمیل پژوهشگر مکاتبه شود.

نتایج جدول ۲ نشان می‌دهد که واژه‌های موجود در متن مقالات پس از اعمال الگوی زو تبدیل به ۳۴۶۸ واژه شدند که از بین آن‌ها قبل از تعدیل واژه‌ها، ۱۳۶۸ واژه و بعد از تعدیل آن‌ها، ۴۶۸ واژه به عنوان واژه بازدارنده استخراج شدند که در جدول فقط ۱۰واژه اول ارائه شده است.

۲. واژه‌های بازدارنده موجود در چکیده مقاله‌ها جهت به کارگیری در نمایه‌سازی خودکار مدارک فارسی رشته کتابداری و اطلاع‌رسانی بر اساس الگوی زو، کدام است؟

جدول ۳. فهرست واژه‌های بازدارنده چکیده مقاله‌ها بر اساس الگوی زو «قبل و بعد از تعدیل»

ردیف	واژه	قبل از تعدیل	بعد از تعدیل	میانگین احتمال	واریانس احتمال	ارزش آماری واژه
۱	است	✓	✓	۰,۰۱۰۵۵	۰,۰۰۰۰۳۳۷	۵۵,۰۴
۲	استفاده	✓	–	۰,۰۰۵۷۸	۰,۰۰۰۰۶۱۸	۱۳۶,۱۳
۳	استفاده شد	✓	–	۰,۰۰۱۵۳۵	۰,۰۰۰۰۰۶۴۱	۱۶۴,۹۲
۴	افزایش	✓	–	۰,۰۰۱۶۰۵	۰,۰۰۰۰۰۶۷۷	۱۶۲,۱۱
۵	اطلاعاتی	✓	–	۰,۰۰۲۸۸	۰,۰۰۰۰۲۴۹	۱۷۳,۵۹
۶	ارائه	✓	✓	۰,۰۰۲۲۲	۰,۰۰۰۰۲۳۳	۲۱۷,۹۱
۷	ارائه شده	✓	✓	۰,۰۰۲۲۲	۰,۰۰۰۰۲۳۳	۲۱۷,۹۱
۸	از	✓	✓	۰,۰۲۳۵۷	۰,۰۰۰۱۵۸	۵۳,۳۵
۹	از نظر	✓	✓	۰,۰۰۲۰۳	۰,۰۰۰۰۱۵۳	۱۹۲,۷۵
۱۰	این	✓	✓	۰,۰۱۶۶۴	۰,۰۰۰۱۲۲	۶۶,۴۴
	جمع	۱۴۸	۱۷۳			

نتایج جدول ۳ نشان می‌دهد که واژه‌های موجود در چکیده مقالات پس از اعمال الگوی زو تبدیل به ۹۱۳ واژه شدند که از میان آن‌ها قبل از تعدیل، ۱۴۸ واژه و بعد از تعدیل، ۱۷۳ واژه به عنوان بازدارنده استخراج شدند.

۳. واژه‌های بازدارنده موجود در متن مقاله‌ها جهت به کارگیری در نمایه‌سازی خودکار مدارک فارسی رشته کتابداری و اطلاع‌رسانی بر اساس قانون زیف، کدام است؟

جدول ۴. فهرست واژه‌های بازدارنده متن مقاله‌ها بر اساس الگوی زیف«قبل و بعد از تعدیل»

	قبل از تعدیل			بعد از تعدیل		
ردیف	واژه	بسامد	تعداد مقالاتی که عامل انتخاب واژه بوده‌اند	واژه	بسامد	تعداد مقالاتی که این واژه را در بر دارند
۱	ابرداده	۲۹	۱	احتمالاً	۳	۳
۲	اتخاذ	۲۹	۱	اساساً	۱	۱
۳	اثر	۲۱	۱	ایشان	۲	۱
۴	اختلاف	۲۳	۱	اینجا	۸	۵
۵	ارائه	۲۴	۱	اینک	۱	۱
۶	ارتباطات	۳۰	۱	اغلب	۲۱	۱۱
۷	از	۱۵۳۷	۲۰	✓	۱۵۳۷	۲۰
۸	استناد	۱۰۰	۱	افزون بر این	۳	۱
۹	است	۵۷۵	۱۴	✓	۶۵۷	۲۰
۱۰	استاندارد	۲۶	۱	اکثر	۲۰	۲۰

نتایج جدول ۴ نشان‌دهنده آن است که از بین ۱۵۵۵۷ واژه مورد مطالعه، مطابق قانون زیف و قبل از تعدیل، ۲۱۷ واژه و بعد از تعدیل، ۶۰۷ واژه به عنوان واژه بازدارنده شناخته شدند.

۴. واژه‌های بازدارنده موجود در چکیده مقاله‌ها جهت به کارگیری در نمایه‌سازی خودکار مدارک فارسی رشته کتابداری و اطلاع‌رسانی بر اساس قانون زیف، کدام است؟

جدول۵. فهرست واژه‌های بازدارنده موجود در چکیده‌ها بر اساس قانون زیف «قبل و بعد از تعدیل»

	قبل از تعدیل			بعد از تعدیل		
ردیف	واژه	بسامد	تعداد مقالاتی که عامل انتخاب واژه بوده‌اند	واژه	بسامد	تعداد مقالاتی که این واژه‌را در بر دارند
۱	اثر	۴	۱	ارائه	۶	۴
۲	از	۷۳	۱۱	✓	۸۷	۱۸
۳	استفاده	۱۲	۲	ازاین‌رو	۱	۱
۴	استناد	۸	۱	از آن‌جا که	۱	۱
۵	اطلاعات	۷	۱	از جمله	۲	۲
۶	اعضای	۱۴	۲	از لحاظ	۴	۲
۷	این	۳۸	۶	✓	۶۰	۱۵
۸	آن	۱۲	۲	این‌گونه	۱	۱
۹	بُعد	۵	۱	این که	۲	۲
۱۰	با	۳۵	۶	✓	۶۱	۱۷

همان گونه که در جدول ۵ مشاهده می‌شود از میان ۱۹۸۹ واژه موجود در چکیده‌های مورد مطالعه قبل از تعدیل، ۶۰ واژه و بعد از تعدیل، ۱۸۶ واژه به عنوان واژه بازدارنده شناسایی شدند.

۵. تا چه حد بین سیاهه‌های واژه‌های بازدارنده به دست آمده از متن و چکیده، در روش زو هم‌خوانی وجود دارد؟

قبل از تعدیل، از بین ۱۳۷۵ واژه مورد مقایسه،۱۴۱ واژه (۱۰/۲۵ درصد) و بعد از تعدیل از بین ۴۸۸ واژه مورد مقایسه، ۱۵۳ واژه (۳۱/۳۵درصد) بین دو فهرست حاصل از متن و چکیده با هم مشترک هستند.

۶. تا چه حد بین سیاهه‌های واژه‌های بازدارنده به دست آمده از متن و چکیده، در روش زیف هم‌خوانی وجود دارد؟

از بین ۲۱۸ واژه مورد مقایسه قبل از تعدیل، ۵۹ واژه بازدارنده مشترک (۲۷/۰۶ درصد) بین دو فهرست مورد اشاره یافت شد. از ۶۱۳ واژه مورد مقایسه پس از تعدیل واژه‌ها، ۱۸۰ واژه بازدارنده (۲۹/۳۶ درصد) بین دو فهرست مشترک بودند.

۷. تا چه حد بین سیاهه‌های واژه‌های بازدارنده به دست آمده از متن مقاله‌های دو روش زو و زیف هم‌خوانی وجود دارد؟

قبل از تعدیل، از مجموع ۱۴۵۶ واژه مورد مقایسه، ۱۲۹ واژه بازدارنده (۰۸/۸۵ درصد) و پس از تعدیل واژه‌ها نیز از میان ۶۶۴ واژه مورد مقایسه، ۴۱۱ واژه (۶۱/۸۹ درصد) بین دو فهرست مشترک بودند.

۸. تا چه حد بین سیاهه‌های واژه‌های بازدارنده به دست آمده از چکیده مقاله‌های دو روش زو و زیف هم‌خوانی وجود دارد؟

قبل از تعدیل، ۱۶/۲۰ درصد (یعنی ۲۹ واژه از ۱۷۹ واژه مورد مقایسه) از واژه‌های بازدارنده و پس از تعدیل نیز، ۸۳/۱۶ درصد (یعنی ۱۶۳ واژه از ۱۹۶ واژه مورد مقایسه) بین این دو فهرست مشترک بودند.

جدول ۶. فهرست واژه‌های بازدارنده مشترک بین فهرست‌های مورد بررسی به تفکیک قبل و بعد از تعدیل

ردیف	۲۰ واژه بازدارنده مشترک بین متن و چکیده زو		۲۰ واژه بازدارنده مشترک بین متن و چکیده زیف		۲۰ واژه بازدارنده مشترک بین متن‌های دو روش زو و زیف		۲۰ واژه بازدارنده مشترک بین چکیده‌های دو روش زو و زیف	
	قبل از تعدیل	بعد از تعدیل	قبل از تعدیل	بعد از تعدیل	قبل از تعدیل	بعد از تعدیل	قبل از تعدیل	بعد از تعدیل
۱	از	ارائه	اثر	ارائه	اثر	احتمالاً	از	از
۲	است	از	از	ارائه‌شده	ارائه	ارائه	استفاده	از این رو
۳	اطلاعاتی	از نظر	استفاده	از	از	ارائه شده	اعضای	از این لحاظ
۴	اعضای	از آن جا که	استناد	ازاین‌رو	از نظر	از	این	از آن جا که
۵	افزایش	از طریق	اطلاعات	از آن‌جا که	از طریق	از این دست	آن	از جمله
۶	از	ازجمله	اعضای	ازجمله	است	از آن روست که	با	از طریق
۷	است	از قبیل	این	از لحاظ	استاندارد	از جمله	برای	از قبیل
۸	اطلاعاتی	از لحاظ	آن	از نظر	استفاده	از طریق	به	از لحاظ
۹	اعضای	است	بُعد	از این لحاظ	اطلاعات	از قبیل	بین	از نظر
۱۰	افزایش	اول	با	از قبیل	اطلاعاتی	از لحاظ	پژوهش	است

۹. واژه‌های بازدارنده چه حجمی از متن و چکیده مقاله‌های حوزه مورد بررسی را تشکیل می‌دهند؟

به منظور پاسخگویی به این سؤال، نسبت تعداد واژه‌های بازدارنده متن یا چکیده به تعداد کل واژه‌های متن/ چکیده به تفکیک الگوی زو و قانون زیف مورد محاسبه قرار گرفت. نتایج بررسی در جدول‌های ۷ و ۸ بیان شده است.

جدول ۷. حجم واژه‌های بازدارنده در متن و چکیده مقالات بر اساس الگوی زو

الگوی زو	نام شیوه	تعداد کل واژه‌های متن بدون احتساب بسامد	تعداد واژه‌های متن بعد از اعمال الگوی زو	تعداد واژه‌های بازدارنده	درصد واژه‌های بازدارنده بعد از اعمال الگوی زو	درصد واژه‌های بازدارنده نسبت به تعداد کل واژه‌های متن بدون احتساب بسامد
متن مقالات	قبل از تعدیل	۱۵۵۵۷	۳۴۶۸	۱۳۶۸	۳۹/۴۴	۰۸/۷۹
	بعد از تعدیل	۱۵۵۵۷	۳۴۶۸	۴۶	۱۳/۴۹	۳
چکیده مقالات	قبل از تعدیل	۱۹۸۹	۹۱۳	۱۴۸	۱۶/۲۱	۰۷/۴۴
	بعد از تعدیل	۱۹۸۹	۹۱۳	۱۷۳	۱۸/۹۴	۰۸/۶۹

یافته‌های جدول ۷ نشان‌دهنده این است که در الگوی زو بالاترین نسبت واژه‌های بازدارنده قبل از تعدیل واژه‌ها، مربوط به متن مقالات (۳۹/۴۴ درصد) و پس از تعدیل مربوط به چکیده مقالات (۱۸/۹۴ درصد) می‌باشد.

جدول ۸. حجم واژه‌های بازدارنده در متن و چکیده مقالات بر اساس فرمول زیف

فرمول زیف		بدون احتساب بسامد			با احتساب بسامد		
	نام شیوه	تعداد کل واژه‌های متن بدون احتساب بسامد	تعداد واژه-های بازدارنده بدون احتساب بسامد	درصد واژه-های بازدارنده بدون احتساب بسامد	تعداد کل واژه‌های متن با احتساب بسامد	تعداد واژه-های بازدارنده با احتساب بسامد	درصد واژه-های بازدارنده با احتساب بسامد
متن مقالات	قبل از تعدیل	۱۵۵۵۷	۲۱۷	۰۱/۳۹	۷۳۷۵۰	۳۵۴۷۷	۴۸/۱۰
	بعد از تعدیل	۱۵۵۵۷	۶۰۷	۰۳/۹	۷۳۷۵۰	۳۲۴۲۸	۴۳/۹۷
چکیده مقالات	قبل از تعدیل	۱۹۸۹	۶۰	۰۳/۰۱	۳۸۲۰	۱۱۳۷	۲۹/۷۶
	بعد از تعدیل	۱۹۸۹	۱۸۶	۰۹/۳۵	۳۸۲۰	۱۶۲۵	۴۲/۵۳

یافته‌های جدول ۸ نشان می‌دهد که در روش زیف، متن مقالات قبل از تعدیل (۴۸/۱۰ درصد) و پس از تعدیل (۴۳/۹۷ درصد) بالاترین نسبت واژه‌های بازدارنده را دارا است.

نتیجه‌گیری

در متن مقالات مطابق الگوی زو تعداد واژه‌های بازدارنده قبل از تعدیل، ۶۵/۷۹ درصد بیش‌تر از بعد از تعدیل می‌باشد. به نظر می‌رسد این تفاوت در تعداد واژه‌های بازدارنده ناشی از وجود تعداد قابل توجهی از واژه‌های معنادار در فهرست بازدارنده‌ای باشد که در شیوه قبل از تعدیل ایجاد شده است، به نحوی که با حذف این واژه‌ها در شیوه بعد از تعدیل، تعداد واژه‌ها در دومین فهرست کاهش یافت. نتایج این پژوهش تا اندازه زیادی با یافته‌های پژوهش سنجی (۱۳۸۷)؛ فاکس (۱۹۹۰)؛ هائو و هائو (۲۰۰۸)؛

داورپناه، سنجی و آرمیده (۲۰۰۹) مطابقت دارد. سنجی، ۱۲۹۱ واژه بازدارنده را بر اساس نوع دستوری واژه‌ها و با در نظر گرفتن بسامد آن‌ها شناسایی نمود. فاکس (۱۹۹۰)؛ هائو و هائو (۲۰۰۸) و داورپناه، سنجی و آرمیده (۲۰۰۹) نیز در پژوهش‌های خویش به ترتیب ۴۲۱، ۵۰۰ و ۴۲۲ واژه بازدارنده را استخراج نمودند. به نظر می‌رسد تفاوت بین تعداد واژه‌های بازدارنده این پژوهش و سایر پژوهش‌های مورد بررسی به دلیل تفاوت در روش مورد استفاده جهت استخراج واژه‌های غیرمفهومی، تفاوت در جامعه آماری مورد مطالعه یا اختلاف بین پژوهشگران در تعریف واژه و تعیین مرز کلمات که یک مرحله اجتناب‌ناپذیر قبل از شناسایی واژه‌های بازدارنده است و یا تفاوت در نقطه برش تعیین شده باشد. این نقطه برش بر اساس استنباط فردی پژوهشگر و بررسی پیکره مورد مطالعه مشخص می‌شود. با توجه به این که در الگوی زو بر اساس محاسبات آماری، ارزش بسامد واژه در کل خوشه زبانی مورد توجه قرار می‌گیرد، می‌توان از این الگو به نحو شایسته‌تری در بانک‌های اطلاعاتی تخصصی و مجموعه متون هم‌بند و هم‌سنخ استفاده نمود زیرا بسیاری از واژه‌های عام و بی‌ارزش محتوایی و اطلاعاتی را در مجموعه‌های بزرگ شناسایی نموده و باعث بهینه‌سازی ذخیره و بازیابی اطلاعات تخصصی می‌شود و به نمایه‌سازان، چکیده‌نویسان و طراحان پایگاه‌ها و نرم‌افزارهای اطلاعاتی در کاهش میزان حجم فایل مقلوب نمایه، زمان و هزینه ذخیره‌سازی و بازیابی اطلاعات کمک می‌نماید.

در چکیده مقالات نیز طبق الگوی زو، تعداد واژه‌های بازدارنده بعد از تعدیل، ۱۴/۴۵ درصد افزایش یافت. به نظر می‌رسد این تفاوت در تعداد واژه‌های بازدارنده، می‌تواند مربوط به نقطه برش تعیین شده برای چکیده مقالات باشد که بر اساس آن برخی از واژه‌های غیرمفهومی به دلیل این که ارزش آماری بالایی داشتند، در بخش واژه‌های معنادار قرار گرفتند و با اجرای شیوه بعد از تعدیل و اضافه نمودن آن‌ها به فهرست، تعداد واژه‌های بازدارنده در دومین فهرست افزایش یافت. تنها پژوهشی که به بررسی چکیده مقالات به منظور شناسایی واژه‌های بازدارنده پرداخته بود، پژوهش ساووی (۱۹۹۹) بود که در پژوهش خود، ۲۱۵ واژه بازدارنده را از متن و چکیده مقالات استخراج نمود.

بنا بر قانون زیف نیز در متن مقالات قبل از تعدیل، ۲۱۷ واژه (با احتساب بسامد ۳۵۴۷۷ واژه) و بعد از تعدیل، ۶۰۷ واژه (با احتساب بسامد ۳۲۴۲۸ واژه) به عنوان واژه بازدارنده شناخته شدند. همان گونه که مشاهده می‌شود بدون احتساب بسامد، تعداد واژه‌های بازدارنده در شیوه بعد از تعدیل که به مفهوم معنایی کلمات و تعدیل واژه‌ها توجه شده، ۳ برابر شده است اما با احتساب بسامد، تعداد واژه‌های بازدارنده شیوه قبل از تعدیل بیش‌تر است. قانون زیف فقط بسامد واژه‌ها را در نظر می‌گیرد و واژه‌هایی که زیاد تکرار

شده باشند را به فهرست اضافه می‌نماید که البته تعدادی از آن‌ها معنادار بوده و به همین دلیل در شیوه بعد از تعدیل از فهرست واژه‌های بازدارنده حذف شدند؛ در مقابل واژه‌های غیرمعناداری هم که به فهرست شیوه بعد از تعدیل اضافه شدند از بسامد بالایی برخوردار نبودند. یافته‌ها با نتیجه پژوهش داورپناه و بلندیان (۱۳۸۵) مطابقت داشت. آن‌ها به این نتیجه رسیدند که واژه‌های بازدارنده بالاترین بسامد را به خود اختصاص داده‌اند که با حذف آن‌ها از حجم مدارک به میزان قابل توجهی کاسته می‌شود.

در چکیده مقالات نیز مطابق قانون زیف، قبل از تعدیل ۶۰ واژه (با احتساب بسامد آن‌ها، ۱۱۳۷ واژه) و بعد از تعدیل، ۱۸۶ واژه (با احتساب بسامد آن‌ها، ۱۶۲۵ واژه) به عنوان واژه بازدارنده شناخته شدند که با اجرای شیوه بعد از تعدیل، ۷۴/۸۴ درصد افزایش داشته است. تعداد کم واژه‌های بازدارنده استخراج شده از چکیده مقالات نشان می‌دهد که در سال‌های اخیر پژوهشگران در نگارش چکیده‌های خود دقت بیش‌تری می‌کنند و چکیده‌ها نیز بار اطلاعاتی افزون‌تری را منتقل کرده و بیشتر حاوی واژه‌های مفهومی می باشند.

الگوی زو یکی از الگوهای آماری است که برای شناسایی واژه‌های بازدارنده مورد استفاده قرار گرفته است و بر اساس بسامد عمل می‌کند و ارتباط آن با قانون زیف نیز این است که هر دو مبتنی بر بسامد هستند. در این پژوهش این مسئله مدنظر قرار گرفت و سعی شد قابلیت به کارگیری آن‌ها در زبان فارسی و تفاوت آن‌ها در ایجاد فهرست واژه‌های بازدارنده مورد بررسی قرار گیرد که به طور کلی به نظر می‌رسد که می‌توان در زبان فارسی از الگوی آماری زو و هم‌چنین قانون زیف برای شناسایی واژه‌های بازدارنده موجود در متن و چکیده مقالات استفاده نمود که هر کدام از این دو الگو نکات قوت خاص خود را داراست. البته تشخیص کاربردپذیری و میزان کارایی این فهرست‌ها نیازمند به کارگیری آنان در شرایط عملیاتی تجربی و آزمایشی در حیطه ذخیره و بازیابی متون اطلاعاتی در شاخه‌های موضوعی مختلف می‌باشد.

تعداد واژه‌های بازدارنده موجود در متن بیشتر از چکیده می‌باشد زیرا از یک سو، متن از حجم واژگانی بیشتری نسبت به چکیده برخوردار است و از سوی دیگر، چکیده معمولاً غنی‌تر از متن است و بیشتر حاوی واژه‌های مفهومی می‌باشد. لذا استنباط می‌شود که طول متن در حجم واژه‌های بازدارنده تأثیر دارد.

در هر دو روش رابطه مستقیمی بین بسامد واژه‌ها (در متن و چکیده) و احتمال بازدارنده بودن آن‌ها یافت شد. هم‌چنین عامل تعدیل که در این پژوهش مورد استفاده قرار گرفت نقش بسزایی در شناسایی

واژه‌های بازدارنده داشت که در آن، حذف بعضی از واژه‌ها و اضافه نمودن برخی دیگر بر اساس قضاوت پژوهشگر انجام گرفته و نیاز به آگاهی و مهارت وی در مسائل زبان مورد مطالعه دارد. اما در مقایسه با شیوه قبل از تعدیل برای استخراج واژه‌های بازدارنده زمان و هزینه‌های بیشتری را می‌طلبد. در کلیه فهرست‌های به دست آمده از پژوهش میزان هم‌خوانی با اجرای شیوه دوم افزایش یافته است. در این شیوه عامل انسانی دخالت داشته و از معیارهای مشابهی برای تمایز واژه‌های معنادار و غیرمعنادار استفاده شده است، در نتیجه تعداد واژه‌های مشترک بین فهرست‌های مورد بررسی بیشتر است. در مجموع به قصد مقایسه و کاربرد این دو روش در استخراج واژه های بازدارنده این پژوهش انجام شد و با استفاده از دو روش مذکور، ۸ فهرست از واژه‌های بازدارنده موجود در متن و چکیده مقالات تهیه شد و میزان هم‌خوانی بین آن‌ها به تفکیک دو شیوه، مورد مطالعه قرار گرفت که یافته‌ها نشان داد واژه‌های موجود در این فهرست‌ها تا حد زیادی متشابه نبوده و میزان هم‌خوانی بین این فهرست‌ها در سطح پایینی است. این امر نشان‌دهنده آن است که ایجاد فهرست واژه‌های بازدارنده منسجم و واحد، حتی برای یک حیطه موضوعی خاص کاری بس مشکل و تقریباً ناممکن است و نمی‌توان به سادگی به آن دست یافت و به همین دلیل پژوهشگران از روش‌های گوناگونی در این زمینه استفاده کرده‌اند.

بدون احتساب بسامد متن مقالات در روش زو (۳۹/۴۴ درصد) و با احتساب بسامد، متن مقالات در روش زیف (۴۸/۱۰ درصد) دارای بیشترین نسبت واژه‌های غیرمفهومی می‌باشند. با بررسی فهرست به دست آمده از طریق قانون زیف مشخص شد که استفاده از این قانون به عنوان مبنایی برای تعیین کلمات معنادار در یک مدرک مناسب‌تر باشد و یا برای تهیه فهرست واژه‌های بازدارنده در یک مجموعه و پیکره عمومی کاربرد مناسب‌تری داشته باشد که در نهایت، بررسی‌های بیشتر به منظور روشن‌تر شدن ابعاد مختلف یافته‌های این پژوهش از جنبه‌های مختلف نظری و کاربردی، ضروری به نظر می‌رسد.

با توجه به نتایج به دست آمده، پیشنهادهای اجرایی زیر مؤثر به نظر می‌رسد:

۱. با توجه به شیوه خاص نگارش زبان فارسی (فاصله بین کلمات، املای متفاوت واژه‌ها و غیره) که مشکلاتی را در تفکیک واژه‌ها به وجود می‌آورند، پیشنهاد می‌شود دستنامه واحد و استانداردی برای تایپ فارسی و ورود اطلاعات به رایانه مدون گردد تا کلمات به روشنی قابل تفکیک و تشخیص باشند.

۲. از آن جایی که نشریات علمی و پژوهشی دارای بیشترین میزان کاربرد در عملیات ذخیره و بازیابی اطلاعات می‌باشند، تهیه شیوه‌نامه‌ای استاندارد در طول نگارش این متون می‌تواند برای

درونداد اطلاعات در پایگاه‌های اطلاعاتی و پژوهش‌هایی از این دست که با تحلیل واژگان سروکار دارند، مفید باشد.

۳. پیشنهاد می‌شود که هنگام ورود اطلاعات متنی به رایانه در به کارگیری و استفاده از نیم‌فاصله دقت کافی انجام پذیرد به ویژه در مورد نام کامل نویسندگان، افعال و علائم جمع و تکواژهای صرفی و سایر واژه‌هایی که به لحاظ مفهومی قابل تفکیک نمی‌باشند و در همین راستا با توجه به کاربرد متوالی نیم‌فاصله در تایپ فارسی به طراحان سخت‌افزار پیشنهاد می‌شود صفحه کلید تایپ فارسی را به گونه‌ای باز طراحی نمایند که کلیدی جداگانه و سهل‌الوصول برای نیم‌فاصله به صورت مجزا تعبیه گردد.

۴. به طراحان پایگاه‌های اطلاعاتی پیشنهاد می‌شود که با شناسایی واژه‌های بازدارنده و ذخیره آن‌ها در پایگاه خود، میزان کارایی در ذخیره و بازیابی مدارک را افزایش دهند.

۵. پیشنهاد می‌شود در نرم‌افزارهای کتابخانه‌ای و ذخیره و بازیابی اطلاعات، بانک واژه‌های بازدارنده تعبیه شود تا ضمن کاهش مدت زمان و هزینه نمایه‌سازی، مدارک مرتبط بیشتری بازیابی شود.

۶. پیشنهاد می‌شود فهرست‌های به دست آمده از این پژوهش در خوشه‌های واژگانی مختلف عملاً به کار گرفته شود و میزان تأثیر آن در عملیات ذخیره و بازیابی اطلاعات از جمله ضریب دقت و بازیافت، صرفه‌جویی در زمان، هزینه و مانند آن مورد بررسی قرار گیرد.

کتابنامه:

بلندیان، صدیقه (۱۳۸۵). تحلیل متن مقالات فارسی کتابداری و اطلاع‌رسانی و امکان نمایه سازی ماشینی آن‌ها بر اساس قانون زیف. پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد.

تیلور، آرلین(۱۳۸۱). *سازماندهی اطلاعات*. (محمد حسین دیانی، مترجم). مشهد: کتابخانه رایانه‌ای .

سنجی، مجیده (۱۳۸۷). *شناسایی واژه‌های غیرمفهومی رایج در نمایه‌سازی خودکار مدارک فارسی*. پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد.

گیلوری، عباس(۱۳۷۹). نمایه سازی خودکار: گذشته، حال، آینده. *پیام کتابخانه*، ۱۰(۴): ۱۷-۲۵.

هویدا، علیرضا (۱۳۷۸). *آمار و روش‌های کمّی در کتابداری و اطلاع‌رسانی*. تهران: سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها (سمت).

Abu-El Khair, I. H. (2003). Effects of Stop Words Elimination for Arabic Information Retrieval. International Journal of Computing & Information Science, 4(3), 119-133. Retrieved June 18, 2010, from http://www.mons.edu.eg . pcvs/13702/13102.asp

Berg, C. N. (1997). *Developing Corpus Specific Stop Word List Using Quantitative Comparison*. PhD thesis, Graduate school of Logistics and acquisition management, Retrieved November 20, 2010, from http://www.research.airuniv .edu/papers/ay1997/afit/ berg cn.pdf

Davarpanah, M. R., Sanji, M., & Aramideh, M. (2009). Farsi Lexical Analysis and Stop Word List. *Library Hi Tech*, 27(3), 435-449. Retrieved December 14, 2011 from http://www.emeraldinsight.comType=Article&contentId=1811864 /InsightviewContentItem/do?content

Edmundson, H. P., & Wyllys, R. E. (1959). *Automatic Indexing and Abstracting of Contents of Documents*. Retrieved June 14 , 2011, from http:// www. washington.edu

Fox, C. (1990). *A stop list for general text*. Retrieved November 20, 2010, from http://www.informatik.uni-trier.de/ley/indice/a-tree.pdf

Hao, L., & Hao, Li. (2008). Automatic Identification of Stop Words in Chinese Text Classification. Retrieved October 3, 2011, from http://ieeexplore .ieee.org/xpls,/abs_all.jsp? arnumber =4721858

Kerner, Y. H., & Blitz, S.Y. (2010). *Experiments With Extraction* of *Stop words in Hebrew*. Retrieved April 21, 2012, from *http://www. cs.tau.a*c.ir/~nachum/iscol/HaCohenKerner_ISCOL 10_2.pdf

Lazarinis, F. (2007). Engineering and Utilizing a Stop Word List in Greek Web. *Journal of the American Society for Information Science and Technology*, 58(11), 1645-1652. Retrieved November 18, 2011, from http://dl.acm.org/ citation.cfm?id=1285331

Pandey, A. K., & Siddiqui, T. (2009). *Evaluation Effect of Stemming and Stop- Word Removal on Hindi Text Retrieval.* Retrieved September 17, 2010, from http://www.springerlink.com/index/j6444068.x213572k. pdf Savoy, J. (1999). A Stemming Procedure and Stop Word List for General French Corpora. *Journal of the American Society for Information Science*, 50(10), 944-952. Retrieved September 17, 2010, from http://www. members.unine.ch /jacques.savoy/papers/frjasis.pdf

Wilbur, j., & Sirotkn, K (1992). The automatic identification of Stop Word. *Journal of Information Science*, 18 (1), 45-55. Retrieved September 3, 2010, from http://www. jis.sagepub.com/content/18/1/4

zou, F., Deng, X., & Han, S. (2006). *Automatic identification of Chinese Stop Words*. Retrieved November 10, 2010 , from http:/ www.cicling. org/2006/ RCS-18/RCS-18-Page151.pdf